18

THE GEOMETRIC MEAN IN VITAL AND SOCIAL STATISTICS.

BY

FRANCIS GALTON, F.R.S.

THE LAW OF THE GEOMETRIC MEAN.

BY

DONALD McALISTER, B.A., B.Sc.,
Fellow of St. John's College, Cambridge.

6 MAR 23
LIBRARY

[*From the* PROCEEDINGS OF THE ROYAL SOCIETY, No. 198, 1879.]

"The Geometric Mean, in Vital and Social Statistics." By FRANCIS GALTON, F.R.S. Received October 21, 1879.

My purpose is to show that an assumption which lies at the basis of the well-known law of "Frequency of Error" (commonly expressed by the formula $y=e^{-h^2x^2}$), is incorrect in many groups of vital and social phenomena, although that law has been applied to them by statisticians with partial success and corresponding convenience. Next, I will point out the correct hypothesis upon which a Law of Error suitable to these cases ought to be calculated; and subsequently I will communicate a memoir by Mr. Donald McAlister, who, at my suggestion, has mathematically investigated the subject.

The assumption to which I refer is, that errors in excess or in deficiency of the truth are equally probable; or conversely, that if two fallible measurements have been made of the same object, their arithmetical mean is more likely to be the true measurement than any other quantity that can be named.

This assumption cannot be justified in vital phenomena. For example, suppose we endeavour to match a tint; Fechner's law, in its approximative and simplest form of sensation=log stimulus, tells us that a series of tints, in which the quantities of white scattered on a black ground are as 1, 2, 4, 8, 16, 32, &c., will appear to the eye to be separated by equal intervals of tint. Therefore, in matching a grey that contains 8 portions of white, we are just as likely to err by selecting one that has 16 portions as one that has 4 portions. In the first case there would be an error in excess, of 8; in the second there would be an error in deficiency, of 4. Therefore, an error of the same magnitude in excess or in deficiency is not equally probable in the judgment of tints by the eye. Conversely, if two persons, who are equally good judges, describe their impressions of a certain tint, and one says that it contains 4 portions of white and the other that it contains 16 portions, the most reasonable conclusion is that it really contains 8 portions. The arithmetic mean of the estimates is $\frac{4+16}{2}$ or 10, which is not the most probable value. It is the geometric mean 8 (4 : 8 : : 8 : 16) which is the most probable.

Precisely the same condition characterises every determination by any of the senses; for example, in judging of the weight of bodies and of their temperatures, of the loudness and of the pitch of tones, and of estimates of lengths and distances *as wholes*. Thus, three rods of the lengths a, b, c, when taken successively in the hand, appear to differ by equal intervals when $a : b :: b : c$, and not when $a-b=b-c$. In

all physiological phenomena, where there is on the one hand a stimulus and on the other a response to that stimulus, Fechner's law may be assumed to prevail; in other words, the true mean is the geometric.

The same condition of the geometric mean appears to characterise the majority of the influences, which, combined with those of purely vital phenomena, give rise to the events with which sociology deals. It is difficult to find terms sufficiently general to apply to the varied topics of sociology, but there are two categories of causes, which are of common occurrence. The one is that of ordinary increase, as exemplified by the growth of population, where an already large nation tends to become larger than a small one under similar circumstances, or when a capital employed in a business increases in proportion to its size. The other category is that of surrounding influences, or "milieux" as they are often called, such as a period of plenty in which a larger field or a larger business yields a greater excess over its mean yield than a smaller one. Most of the causes of those differences with which sociology are concerned, and which are not purely vital phenomena, such as those already discussed, may be classified under one or other of these two categories, or under such as are in principle almost the same. In short, sociological phenomena, like vital phenomena are, as a general rule, subject to the condition of the geometric mean.

The ordinary law of Frequency of Error, based on the arithmetic mean, corresponds, no doubt, sufficiently well with the observed facts of vital and social phenomena, to be very serviceable to statisticians, but it is far from satisfying their wants, and it may lead to absurdity when applied to wide deviations. It asserts that deviations in excess must be balanced by deviations of equal magnitude in deficiency; therefore, if the former be greater than the mean itself, the latter must be less than zero, that is, must be negative. This is an impossibility in many cases, to which the law is nevertheless applied by statisticians with no small success, so long as they are content to confine its application within a narrow range of deviation. Thus, in respect of stature, the law is very correct in respect to ordinary measurements, although it asserts that the existence of giants, whose height is more than double the mean height of their race, implies the possibility of the existence of dwarfs, whose stature is less than nothing at all.

It is, therefore, an object not only of theoretical interest but of practical use, to thoroughly investigate a Law of Error, based on the geometric mean, even though some of the expected results may perhaps be apparent at first sight. With this view I placed the foregoing remarks in Mr. Donald McAlister's hands, who contributes the following memoir.

6- 23

"The Law of the Geometric Mean." By DONALD MCALISTER, B.A., B.Sc., Fellow of St. John's College, Cambridge. Communicated by FRANCIS GALTON, F.R.S. Received October 21, 1879.

Suppose we have before us a large number of measurements. They may either be all approximations to the true value of a single unknown quantity, or may refer to the several members of a large class. The measurements will disagree among themselves, but on arranging them in order of size they show a tendency to cluster round some *medium* value. We are naturally inclined to infer that the true value of the unknown, or typical member of the class, is not far from this value. How to define and determine the appropriate *medium* in various classes of measurement becomes thus a natural object of inquiry. On examination we find that there is no strict and final criterion applicable to all cases. We have to start with an empirical assumption, more or less justifiable on general grounds, but not capable of rigid proof. In the ordinary Theory of Errors, which deals primarily with discrepant *observations*, the assumption made is reducible to this :—that, on the whole, the best *medium* we can take is the Arithmetic Mean of the discordant measures. This is equivalent to the statement that errors (or *differences* from the truth) of equal amount in excess or defect are equally likely to occur. In the class of measures referred to by Mr. Galton, which are *not* of the nature of instrumental observations, reason is given for thinking that, on the whole, a better *medium* value is likely to be obtained by taking the Geometric Mean of the discrepant measures. This is equivalent to the assumption that measures are equally likely which bear to the truth

ratios of equal amount in excess or defect, so to speak. For example, that measures which are respectively $(\frac{1001}{1000})$th and $(\frac{1000}{1001})$th of the truth are equally likely to occur. This paper seeks to develop some consequences of this fundamental principle.

The practical outcome is a method for the treatment of a series of measures which naturally group themselves round their Geometric Mean. This method may be briefly presented as follows:—

Let the measures be

$$x_1,\ x_2,\ x_3\ \ .\ \ .\ \ .\ \ .\ \ x_n.$$

Take the (hyperbolic) logarithm of each measure, thus forming a new series, say

$$y_1,\ y_2,\ y_3\ \ .\ \ .\ \ .\ \ y_n;\ \text{where } y_r = \log_e x_r.$$

Form the Arithmetic Mean (A.M.) of the x's, so that

$$n(\text{A.M.}) = \Sigma x.$$

Form the Arithmetic Mean of the y's: this will be the logarithm of the Geometric Mean (G.M.) of the x's, so that

$$n \log(\text{G.M.}) = \Sigma y.$$

If then $\frac{1}{h^2} \equiv 4(\log(\text{A.M.}) - \log(\text{G.M.}))$, we have in h a "measure of precision" of the series of x's which we call the "*weight.*"

The series of y*'s may then be treated as a series grouped round its Arithmetic Mean, and the formulæ of the ordinary theory are at once applicable to it, the weight ascribed to it being* h, *as above defined.*

When the y-series is so treated, we can at once pass from it to the x-series. Thus the probability, frequency, &c., of any x are the same as for the corresponding y: and any term interpolated in the y-series will be the logarithm of the corresponding term in the x-series.

1. If it be granted that the Geometric Mean of the measures is the *most* probable value of the quantity measured, it follows that measures which differ from the mean have a probability which becomes less as the discrepancy becomes greater, either in excess or defect. We naturally seek for some scale which shall define the respective probability to be assigned to each measure of any given series.

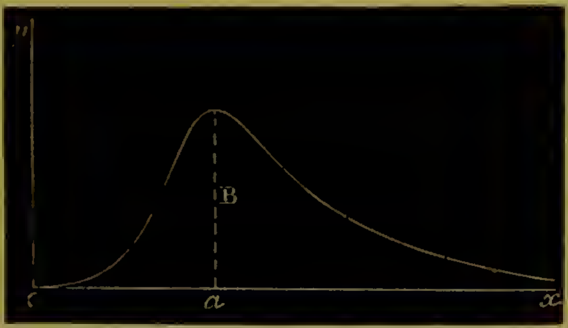
Thus, if x be one of the measures, and a be the Geometric Mean of the whole series, and if we can assign a form to the function ϕ, such that $\phi\left(\frac{x}{a}\right)$ is the probability of x, the curve $y = \phi\left(\frac{x}{a}\right)$ will afford us (graphically) the required scale, and $\phi\left(\frac{x}{a}\right)$ will be the "*Law of Frequency.*"

2. By a method analogous to the first method of Gauss ("Theoria Motus," ii, 3, 186), it can be shown that

$$\phi\left(\frac{x}{a}\right) \equiv \mathrm{B} \exp.\left(-h^2 \log^2 \frac{x}{a}\right) \equiv \mathrm{B}e^{-h^2\left(\log \frac{x}{a}\right)^2},$$

where B and h are constants. B is determined from the condition that the sum of the probabilities of all the measures must be unity; h depends on the grouping of the measures, being large when they agree closely and small when they are generally discordant; h is thus taken as a measure of the precision of the series in general, and is briefly called the "*weight*" *of the series*, or sometimes the "*weight*" *of the mean* derived therefrom. (See fig. 1.)

Fig. 1.



$$y = \mathrm{B} \exp.\left(-h^2 \log^2 \frac{x}{a}\right).$$

3. In forming a scale of probability, $\phi\left(\frac{x}{a}\right)$, which shall be applicable to any series of measures of which the mean and the weight are known, we must suppose the value of ϕ to be given (either graphically or by table) for all possible values of x. In other words, we must suppose the number of entries to be indefinitely great and indefinitely close together. It is clear that the actual probability of any individual measure becomes in such a series infinitesimal, though the relative magnitudes of any pair of probabilities remain unaltered. This is indicated in the formula of §2 by the constant B becoming infinitesimal. To avoid this inconvenience we are led to modify our original problem, and ask, not "what is the probability of a measure x?" but "what is the probability of a measure lying between the very close limits x and $x+\delta x$?"

4. From the principle indicated above, viz., that on the whole there

are as many measures less than the mean as there are greater than the mean, we deduce that the required probability is

$$\frac{h}{\sqrt{\pi}} \exp.\left(-h^2 \log^2 \frac{x}{a}\right) \frac{\delta x}{x}.$$

the constant factor $h \div \sqrt{\pi}$ results from the condition that the sum of all the values of this function from $x=0$ to $x=\infty$ must be unity. And it will be found that

$$\int_0^\infty \frac{h}{\sqrt{\pi}} \exp.\left(-h^2 \log^2 \frac{x}{a}\right) \frac{dx}{x} = 1.$$

as should be the case.

The coefficient of δx, viz. :—$h \exp.\left(-h^2 \log^2 \frac{x}{a}\right) \div \sqrt{\pi} \,.\, x$, I have ventured to distinguish from the previous function by calling it the "*Law of Facility*." This function may be defined in words as "the ratio which the chance that a measure lies within a given small interval bears to the magnitude of the interval.

5. Tracing the curve which represents the "Law of Facility," we find that its form resembles that of the curve of frequency, but that the maximum ordinate occurs at $x=a \exp.\left(-\frac{1}{2h^2}\right)$ in the former, and of course at $x=a$ in the latter. The latter result implies that the mean is the most probable value: the former means that there is more chance of a measure lying within a given small fraction (δx) of $a \,.\, \exp.\left(-\frac{1}{2h^2}\right)$ than within the same small fraction of any other value that can be named.

6. We may graphically represent our results in three ways :—

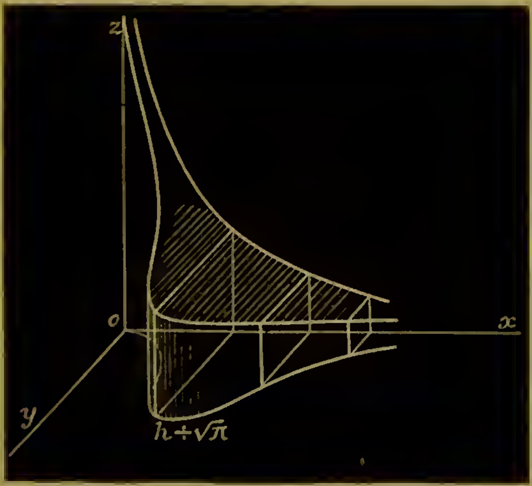
1°. Trace the curve $y\sqrt{\pi}=h \exp.\ (-h^2 \log^2 x)$, (the mean being taken as unity). Let the space between the curve and Ox be the boundary of a lamina of varying density. Take the density along the ordinate at x to be $\frac{1}{x}$. Then the ordinate at x is proportional to the (infinitesimal) probability of x, and the mass of the section bounded by the ordinates at x_1, x_2 measures the probability of a measure lying between x_1 and x_2. If we are dealing with a very large number of measures, this mass will be found to represent the number of measures which actually lie between those limits. The mass of the lamina is bisected by the ordinate at $x=1$ (fig. 2).

FIG. 2.

$$\begin{cases} y\sqrt{\pi} = h \exp.\,(-h^2 \log^2 x). \\ \text{density at } x = \dfrac{1}{x}. \end{cases}$$

2°. Regard the equation $y\sqrt{\pi}=h$ exp. $(-h^2 \log^2 x)$ as that of a cylinder. Let it be cut by another cylinder $zx=1$, at right angles to it. The solid included between these surfaces and the planes xz, xy, obviously corresponds to the lamina of 1°, volume being read for mass (fig. 3). Methods 1° and 2° thus represent in one figure both the frequency and the facility functions.

FIG. 3.



$$\begin{cases} y\sqrt{\pi} = h \exp.\,(-h^2 \log^2 x) \\ z = \dfrac{1}{x}. \end{cases}$$

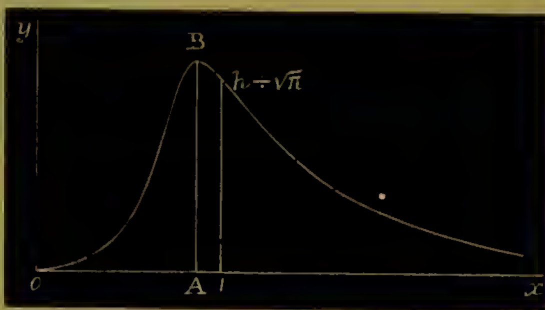
3°. The law of facility alone will be represented by the curve

$$xy\sqrt{\pi}=h \text{ exp. } (-h^2 \log^2 x).$$

Here the area included between any two ordinates, the curve, and the

axis of x, measures the probability of a measure lying between the corresponding abscissæ. The area of the curve is bisected at $x=1$, but the maximum ordinate is at $x=\exp.\left(-\frac{1}{2h^2}\right)$ (fig. 4).

FIG. 4.



$$\begin{cases} xy\sqrt{\pi}=h\exp.\left(-h^2\log^2 x\right). \\ \mathrm{OA}=\exp.\left(-\frac{1}{2h^2}\right). \\ \mathrm{AB}=\frac{h}{\sqrt{\pi}}\exp.\left(\frac{1}{2h^2}\right). \end{cases}$$

7. (Quetelet's method.) Suppose the measures we are considering to be a series of estimates of the magnitude of some single object (such as the depth of a given tint). In forming such an estimate the mind is acted on by many small causes tending to make us err either in excess or defect. Suppose each small cause acts in such a way as to make us under- or over-estimate in a fixed ratio; and as a particular case let this fixed ratio be the same for all the small causes. If, then, half the causes tend to over- and half to under-estimation, we estimate rightly. Our estimate in any given case depends on the particular combination of the causes which has (so far as our knowledge is concerned) fortuitously arisen. Let, then, the number of causes be $2n$, the fixed ratio a. The total number of possible combinations will be 2^{2n}. The resulting possible ratios are

$$\alpha^{2n},\ \alpha^{2n-2},\ \ldots\ \alpha^2,\ 1,\ \alpha^{-2},\ \ldots\ \alpha^{-2n+2},\ \alpha^{-2n}.$$

The chance of a correct estimate (ratio 1) is by the theory of combinations $\underline{|2n}\div 2^{2n}\,\underline{|n}\,\underline{|n}$: the chance of an estimate bearing a ratio $\alpha^{\pm r}$ to the truth will be $\underline{|2n}\div 2^{n}\,\underline{|n-r}\,\underline{|n+r}$. Let now n become very large, and α very near to unity; while r remains finite. Then, by a known approximation, the chance ultimately takes the form $\frac{1}{\sqrt{n\pi}}\exp.\left(-\frac{r^2}{n}\right)$. If we are asked the chance of a ratio x, we put

$x=\alpha^r$, and thus find $\log x = r \log \alpha$; the chance required is then

$$\frac{1}{\sqrt{n\pi}}\exp.\left(-\frac{1}{n}\left(\frac{\log x}{\log \alpha}\right)^2\right),$$

corresponding to the law of frequency.

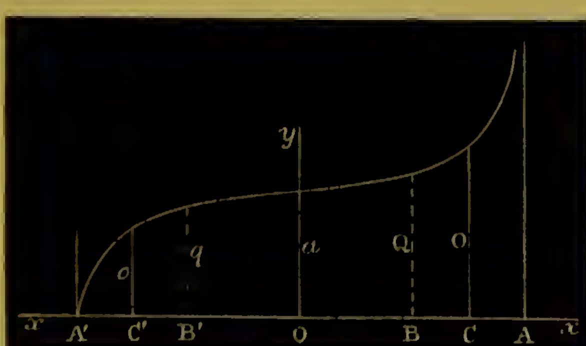
Next determining the chance that the estimate shall lie between narrow limits x and $x+x'$, we find that we have to multiply the above expression by $x' \div x \log \alpha$. Thus, if we take $h^{-2}=n(\log \alpha)^2$, the expression agrees exactly with that already found for the law of facility. This verification is interesting, for though it depends on certain special suppositions, the process seems to throw light on the genesis of the law, and the significance of the modulus h.

8. Another method of exhibiting the law, suggested by Mr. Galton's Method of Intercomparison, is the following. Let the series of measures be represented by a series of ordinates: arrange these side by side at equal small distances and in order of magnitude. Their extremities will then lie on a curve of contrary flexure, which Mr. Galton calls an Ogive; we may speak of it as the "*curve of distribution.*" Its equation can be shown to be

$$hkx=\int_0^{h\log y} e^{-t^2}\,dt \equiv \text{erf.}\,(h\log y),$$

k being a constant depending on the interval between the ordinates as they stand (fig. 5).

FIG. 5.



$$\left\{\begin{aligned} &hkx=\text{erf.}\,(h\log y).\\ &OA=OA'=\sqrt{\pi}\div 2hk.\\ &OB=OB'=\tfrac{1}{2}OA.\\ &BC=B'C'=\tfrac{1}{4}OA.\end{aligned}\right.$$

9. We may next show that if the Arithmetic Mean (A.M.) of all the measures be formed then

$$\frac{1}{h^2}=4\log\left(\frac{\text{A.M.}}{a}\right).$$

This follows strictly when the number of measures is indefinitely great. It still gives a good approximation to the value of h when the number is considerable.

10. Among the measures greater than the mean there is one which may be called middlemost: *i.e.*, such that it is an even wager that a measure (greater than the mean) lies above it or below it. A similar middlemost measure exists among those that are less than the mean. As these two measures, with the mean, divide the curve of facility into four equal parts, I propose to call them the "*higher quartile*" and the "*lower quartile*" respectively. It will be seen that they correspond to the ill-named "probable errors" of the ordinary theory. If Q, q be the quartiles we can show that

$$h\log\frac{Q}{a}=\cdot4769\ldots=-h\log\frac{q}{a},$$

so that $Qq=a^2$. Thus also

$$Q^h(\cdot6208\ldots)=a^h=q^h(1\cdot611\ldots).$$

Similarly between zero and lower quartile we place a mid-measure which we call the "*lower octile*." The "*higher octile*" will subdivide the interval between the higher quartile and infinity. If O, o, be the octiles, we have—

$$O^h(\cdot443)=a^h=o^h(2\cdot255\ldots):$$

as before $Oo=a^2$: and

$$\left(\frac{O}{Q}\right)^h=\left(\frac{q}{o}\right)^h=1\cdot40\ldots$$

It will be observed that in the curve of distribution Q and q are ordinates equidistant from the mean and the terminal ordinates. O is equidistant from Q and the asymptote.

The analogues suggested by the mean error, and mean square error, &c., of the ordinary theory have no very practical value or significance for us. It will be remembered that they are introduced to obviate the difficulties arising from *negative* errors. In our problem these somewhat artificial functions have no special place.*

* The mean-square-measure is $a\exp.\left(\frac{1}{2h^2}\right)$. The geometric mean of the measures greater than mean is $a\exp.\left(\frac{\cdot564\ldots}{h}\right)$.

The gauges of the series given by the weight and the quartiles and octiles will generally suffice.

11. Let two fallible measures be given whose respective weights are known, and let their product be formed. This product will be a fallible measure like its factors. It can be shown that the product is subject to a law of facility of the form already obtained, and further that, as in the ordinary theory, the weight is such that the square of its reciprocal is equal to the sum of the squared reciprocals of the weights of the factors. More generally, if $x_1, x_2 \quad . \quad . \quad .$ be a series of measures whose weights are $h_1, h_2 \quad . \quad . \quad .$, and if z be connected with them by the equation

$$z = x_1^{n_1} \times x_2^{n} \times \ldots,$$

Then h the weight of z is given by

$$\frac{1}{h^2} = \frac{n_1^2}{h_1^2} + \frac{n_2^2}{h_2^2} + \ldots$$

These results are strict; that which follows is closely approximate and applies to series whose weight is not small.

Let $f\,(a_1, a_2 \quad . \quad . \quad .)$ be any function of the fallible measures $a_1, a_2 \quad . \quad . \quad .$, each of which may be the mean of some considerable number of measures, so that the weights are respectively $h_1, h_2 \quad . \quad . \quad .$ Required the weight to be assigned to the value of the function obtained from these fallible values of its variables. We are able to show that if h_f be the weight in question

$$\frac{1}{h_f^2} = \left(\frac{\partial \log f}{\partial \log a_1}\right)^2 \frac{1}{h_1^2} + \left(\frac{\partial \log f}{\partial \log a_2}\right)^2 \frac{1}{h_2^2} + \ldots$$

12. It only remains to establish the practical method presented in § 1. We have

$$\left.\begin{array}{c}\text{Chance of a logarithm} \\ \text{lying between } y \text{ and } y + \partial y\end{array}\right\} = \left\{\begin{array}{c}\text{Chance of a measure} \\ \text{lying between } x \text{ and } x + \partial x\end{array}\right.$$

$$= \frac{h}{\sqrt{\pi}} \exp. \left(-h^2 \log^2 \frac{x}{a}\right) \frac{\partial x}{x} = \frac{h}{\sqrt{\pi}} \exp. \left(-h^2 \overline{y - \log a}^{\,2}\right) \partial y.$$

This is the expression of the ordinary law of grouping of the y's round their arithmetic mean ($\log a$).

It follows that, for example, the number of x's between x_r and x_s must be proportional to

$$\frac{h}{\sqrt{\pi}} \int_{y_r}^{y_s} \exp. \left(-h^2 y^2\right) dy:$$

this can be readily evaluated by means of the ordinary tables of the error-function.

Again, still using h for the weight, the "probable-error" of the y-series corresponds to "quartile" of the x-series. For, by the ordinary theory

$$h\,(\text{probable error}) = \cdot 4769 \ldots$$
$$\therefore = h\,(\log Q - \log a)$$
$$\therefore \log a + \text{prob. error} = \log Q.$$

In words—the term which differs from the Arithmetic Mean by probable error in the y-series is the log. of the quartile in the x-series.

We infer that no new tables are necessary for the practical working of our method. We require only the tables of the error-function and of hyperbolic logarithms. Of course common logarithms may be used, if we remember to introduce the appropriate modulus into our formulæ.

In conclusion, I desire to acknowledge my indebtedness to Mr. Galton, not only for the suggestion of the problem I have here attempted to solve, but also for many valuable practical hints in the working.

HARRISON AND SONS, PRINTERS IN ORDINARY TO HER MAJESTY, ST. MARTIN'S LANE.